Docket No. 0557-4885-2/cb

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

IN RE APPLICATION OF: Atsuo SHIMADA, et al.

GAU: 2772

SERIAL NO: 09/472,249

EXAMINER:

FILED: December 27, 1999

FOR: DOCUMENT PROCESSOR, DOCUMENT CLASSIFICATION DEVICE, DOCUMENT PROCESSING METHOD, DOCUMENT CLASSIFICATION METHOD, AND COMPUTER-READABLE RECORDING MEDIUM FOR RECORDING PROGRAMS FOR EXECUTING THE METHODS ON A COMPUTER

REQUEST FOR PRIORITY

ASSISTANT COMMISSIONER FOR PATENTS
WASHINGTON, D.C. 20231

SIR:

☐ Full benefit of the filing date of U.S. Application Serial Number [US App No], filed [US App Dt], is claimed pursuant to the provisions of **35 U.S.C. §120**.

☐ Full benefit of the filing date of U.S. Provisional Application Serial Number , filed , is claimed pursuant to the provisions of **35 U.S.C. §119(e)**.

■ Applicants claim any right to priority from any earlier filed applications to which they may be entitled pursuant to the provisions of **35 U.S.C. §119**, as noted below.

In the matter of the above-identified application for patent, notice is hereby given that the applicants claim as priority:

COUNTRY	APPLICATION NUMBER	MONTH/DAY/YEAR
JAPAN	10-376576	December 24, 1998
JAPAN	10-369589	December 25, 1998
JAPAN	11-022915	January 29, 1999
JAPAN	11-343890	December 2, 1999

Certified copies of the corresponding Convention Application(s)

■ are submitted herewith

☐ will be submitted prior to payment of the Final Fee

☐ were filed in prior application Serial No. filed

☐ were submitted to the International Bureau in PCT Application Number .
Receipt of the certified copies by the International Bureau in a timely manner under PCT Rule 17.1(a) has been acknowledged as evidenced by the attached PCT/IB/304.

☐ (A) Application Serial No.(s) were filed in prior application Serial No. filed ; and

(B) Application Serial No.(s)

☐ are submitted herewith

☐ will be submitted prior to payment of the Final Fee

Respectfully Submitted,

OBLON, SPIVAK, McCLELLAND,
MAIER & NEUSTADT, P.C.

Marvin J. Spivak
Registration No. 24,913

Surinder Sachar
Registration No. 34,423

Fourth Floor
1755 Jefferson Davis Highway
Arlington, Virginia 22202
Tel. (703) 413-3000
Fax. (703) 413-2220
(OSMMN 11/98)

09/472,249

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: １９９８年１２月２４日

出 願 番 号
Application Number: 平成１０年特許願第３７６５７６号

出 願 人
Applicant (s): 株式会社リコー

１９９９年１０月２２日

特 許 庁 長 官
Commissioner,
Patent Office 近 藤 隆 彦

出証番号 出証特平１１－３０７２８０７

【書類名】　特許願

【整理番号】　9806234

【提出日】　平成10年12月24日

【あて先】　特許庁長官　伊佐山　建志　殿

【国際特許分類】　G06F 15/40

【発明の名称】　文書分類装置および文書分類方法

【請求項の数】　12

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号
株式会社　リコー内

【氏名】　剣持　栄治

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号
株式会社　リコー内

【氏名】　宮地　達生

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号
株式会社　リコー内

【氏名】　嶋田　敦夫

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号
株式会社　リコー内

【氏名】　山崎　真湖人

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号
株式会社　リコー内

【氏名】　武谷　一寿

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号

株式会社　リコー内

【氏名】　長束　哲郎

【特許出願人】

【識別番号】　000006747

【氏名又は名称】　株式会社　リコー

【代表者】　桜井　正光

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【書類名】　　明細書

【発明の名称】　　文書分類装置および文書分類方法

【特許請求の範囲】

【請求項１】　文書の内容に従って文書群を分類する文書分類装置において、文書データ群を入力する文書入力手段と、入力された文書データ群の各文書に対して所定の基準に基づき文書の分割を行い、一つの文書データから一つまたは複数の分割文書データを生成する文書分割手段と、上記文書データと上記分割文書データとの対応を示す文書－分割文書対応マップを生成する文書－分割文書対応マップ生成手段と、上記分割文書データを分類する分割文書分類手段と、上記分割文書分類手段による分類結果に基づいて分割文書分類結果情報を生成する分割文書分類結果生成手段と、上記文書－分割文書対応マップと上記分割文書分類結果情報とを用いて上記文書データの分類結果情報を生成する文書分類結果生成手段とを備えたことを特徴とする文書分類装置。

【請求項２】　請求項１の文書分類装置において、文書データを保存する文書保存手段と、分割文書データを保存する分割文書保存手段と、文書－分割文書対応マップ生成手段により生成された文書－分割文書対応マップを保存する文書－分割文書対応マップ保存手段とを備えたことを特徴とする文書分類装置。

【請求項３】　請求項２の文書分類装置において、分割文書分類結果生成手段により生成された分割文書分類結果情報を保存する分割文書分類結果保存手段を備えたことを特徴とする文書分類装置。

【請求項４】　請求項１、請求項２または請求項３の文書分類装置において、文書分割手段により生成される複数の分割文書データには分割前の文書データそのものを含むことを特徴とする文書分類装置。

【請求項５】　請求項１乃至請求項４の文書分類装置において、文書分割手段が文書データの構造情報を基に文書データを分割する構成にしたことを特徴とする文書分類装置。

【請求項６】　請求項１乃至請求項４の文書分類装置において、文書データに含まれる要素を抽出する文書要素抽出手段と、上記文書要素抽出手段により抽

出された要素に付随する要素付随情報を抽出する要素付随情報抽出手段とを備え、文書分割手段が、上記文書要素抽出手段により抽出された要素、または上記要素と上記要素付随情報抽出手段により抽出された要素付随情報とを用いて上記文書データを分割する構成にしたことを特徴とする文書分類装置。

【請求項７】　請求項１乃至請求項４の文書分類装置において、文書分割手段が、指示された指定範囲に従って文書データの分割を行う構成にしたことを特徴とする文書分類装置。

【請求項８】　請求項１乃至請求項４の文書分類装置において、文書分割手段が、文書データ中の文字数、文数、または文字数と文数の両方を基に文書データを分割する構成にしたことを特徴とする文書分類装置。

【請求項９】　請求項１乃至請求項８の文書分類装置において、文書分類結果生成手段が、文書データを示す情報および上記文書データに付随する代表的情報を、分類結果情報として抽出して提示する構成にしたことを特徴とする文書分類装置。

【請求項10】　請求項９の文書分類装置において、文書分類結果生成手段が、分割文書データを示す情報および上記分割文書データに付随する代表的情報を、分類結果情報として、抽出して提示する構成にしたことを特徴とする文書分類装置。

【請求項11】　文書の内容に従って文書群を分類する文書分類方法において、文書データ群を入力し、入力された文書データ群の各文書に対して所定の基準に基づき文書の分割を行い、一つの文書データから一つまたは複数の分割文書データを生成し、上記文書データと上記分割文書データとの対応を示す文書－分割文書対応マップを生成し、上記分割文書データを分類し、分割文書分類結果情報を生成し、上記文書－分割文書対応マップと上記分割文書分類結果情報とを用いて上記文書データの分類結果情報を生成することを特徴とする文書分類方法。

【請求項12】　請求項11の文書分類方法により文書群の分類を行うプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、入力された文書群を文書の内容に従って分類する文書分類装置に係わり、特に、一つの文書中に複数の話題や意味が含まれていても利用者に理解しやすく分類できる文書分類装置に関する。

【０００２】

【従来の技術】

近年、インターネットなどの普及により、大量の文書群へのアクセスが可能になり、その結果、その文書群を様々な利用者の意図に基づいて、且つ、効率的に利用できるようにする必要性が高まっている。そのため、大量の文書群を意味のあるカテゴリに分類し、文書群の構造を把握するという知的作業が行われ始めている。しかし、このような分類作業を人手により行う場合、その人的及び時間的なコストが膨大なものになるし、また、分類のための知識を分類者のみが有することになるため、分類担当者が代わると分類基準も変わってしまうことになる。

そのため、文書群を人間が分類するような分類基準で自動的に分類しうる文書分類装置が望まれており、例えば、特開平7 －114572号公報に示されているように、文書に含まれる単語から特徴ベクトルを抽出して文書を分類する技術などが提供されるに至っている。しかし、これらの従来技術においては、文書の構成単位を考慮していないため、文書が一つまたは複数の段落記号やタイトルなどにより区切られた構造を持つ場合には、一つの文書の中に複数の話題や意味が含まれてしまい、その結果、利用者がその分類カテゴリを理解し難くなったり、また、ある特定の話題や特定の意味に限定されたカテゴリになったり、利用者の意図するカテゴリとは異なるカテゴリに分類されてしまうという問題が生じている。

なお、特開平6 －176064号公報に示された文脈依存自動分類装置には、文書の段落情報を考慮した文書自動分類を行うことにより分類精度を高めようとするものが開示されているが、本質的に上記の問題を解決するものではない。

【０００３】

【発明が解決しようとする課題】

本発明の課題は、上記の如き従来技術の問題を解決し、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴ

リに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されることがないことにより、利用者がその分類カテゴリを良く理解できる文書分類装置などを提供することにある。

【０００４】

【課題を解決するための手段】

上記の課題を解決するために、請求項１記載の発明では、文書の内容に従って文書群を分類する文書分類装置において、文書データ群を入力する文書入力手段と、入力された文書データ群の各文書に対して所定の基準に基づき文書の分割を行い、一つの文書データから一つまたは複数の分割文書データを生成する文書分割手段と、上記文書データと上記分割文書データとの対応を示す文書－分割文書対応マップを生成する文書－分割文書対応マップ生成手段と、上記分割文書データを分類する分割文書分類手段と、上記分割文書分類手段による分類結果に基づいて分割文書分類結果情報を生成する分割文書分類結果生成手段と、上記文書－分割文書対応マップと上記分割文書分類結果情報とを用いて上記文書データの分類結果情報を生成する文書分類結果生成手段とを備えた。

また、請求項２記載の発明では、請求項１記載の発明において、文書データを保存する文書保存手段と、分割文書データを保存する分割文書保存手段と、文書－分割文書対応マップ生成手段により生成された文書－分割文書対応マップを保存する文書－分割文書対応マップ保存手段とを備えた。

また、請求項３記載の発明では、請求項２記載の発明において、分割文書分類結果生成手段により生成された分割文書分類結果情報を保存する分割文書分類結果保存手段を備えた。

また、請求項４記載の発明では、請求項１、請求項２または請求項３記載の発明において、文書分割手段により生成される複数の分割文書データが分割前の文書データそのものを含む構成にした。

また、請求項５記載の発明では、請求項１乃至請求項４記載の発明において、文書分割手段が文書データの構造情報を基に文書データを分割する構成にした。

【０００５】

また、請求項６記載の発明では、請求項１乃至請求項４記載の発明において、

文書データに含まれる要素を抽出する文書要素抽出手段と、上記文書要素抽出手段により抽出された要素に付随する要素付随情報を抽出する要素付随情報抽出手段とを備え、文書分割手段が、上記文書要素抽出手段により抽出された要素、または上記要素と上記要素付随情報抽出手段により抽出された要素付随情報とを用いて上記文書データを分割する構成にした。

また、請求項７記載の発明では、請求項１乃至請求項４記載の発明において、指示された指定範囲に従って文書分割手段が文書データの分割を行う構成にした。

また、請求項８記載の発明では、請求項１乃至請求項４記載の発明において、文書分割手段が、文書データ中の文字数、文数、または文字数と文数の両方を基に文書データを分割する構成にした。

また、請求項９記載の発明では、請求項１乃至請求項８記載の発明において、文書分類結果生成手段が分類結果情報として、文書データを示す情報および上記文書データに付随する代表的情報を抽出して提示する構成にした。

また、請求項10記載の発明では、請求項９記載の発明において、文書分類結果生成手段が分類結果情報として、さらに、分割文書データを示す情報および上記分割文書データに付随する代表的情報を抽出して提示する構成にした。

また、請求項11記載の発明では、文書の内容に従って文書群を分類する文書分類方法において、文書データ群を入力し、入力された文書データ群の各文書に対して所定の基準に基づき文書の分割を行い、一つの文書データから一つまたは複数の分割文書データを生成し、上記文書データと上記分割文書データとの対応を示す文書－分割文書対応マップを生成し、上記分割文書データを分類し、分割文書分類結果情報を生成し、上記文書－分割文書対応マップと上記分割文書分類結果情報とを用いて上記文書データの分類結果情報を生成する方法にした。

【０００６】

また、請求項12記載の発明では、請求項11の文書分類方法により文書群の分類を行うプログラムをコンピュータ読み取り可能な記録媒体に記録する構成にした。

上記のような手段にしたので、請求項１および請求項11記載の発明では、入力

された文書データ群の各文書が分割され、一つの文書データから一つまたは複数の分割文書データが生成され、上記文書データと上記分割文書データとの対応を示す文書－分割文書対応マップが生成され、上記分割文書データが分類され、分割文書分類結果情報が生成され、上記文書－分割文書対応マップと上記分割文書分類結果情報とを用いて上記文書データの分類結果情報が生成される。

請求項２記載の発明では、請求項１記載の発明において、文書データ、分割文書データ、および文書－分割文書対応マップが保存される。

請求項３記載の発明では、請求項２記載の発明において、さらに、分割文書分類結果情報が保存される。

請求項４記載の発明では、請求項１、請求項２または請求項３記載の発明において、複数の分割文書データのなかに分割前の文書データそのものが含まれる。

請求項５記載の発明では、請求項１乃至請求項４記載の発明において、文書データの構造情報を基にして文書データが分割される。

請求項６記載の発明では、請求項１乃至請求項４記載の発明において、分割対象の文書データから抽出された要素、または上記要素と上記要素から抽出された要素付随情報とを用いて上記文書データが分割される。

請求項７記載の発明では、請求項１乃至請求項４記載の発明において、利用者により指示された指定範囲に従って文書データが分割される。

請求項８記載の発明では、請求項１乃至請求項４記載の発明において、文書データ中の文字数、文数、または文字数と文数の両方を基に文書データが分割される。

請求項９記載の発明では、請求項１乃至請求項８記載の発明において、分類結果情報として、文書データを示す情報および上記文書データに付随する代表的情報が抽出・提示される。

請求項10記載の発明では、請求項９記載の発明において、分類結果情報として、さらに、分割文書データを示す情報および上記分割文書データに付随する代表的情報が抽出・提示される。

請求項12記載の発明では、請求項11記載の発明により文書群の分類を行うプログラムをコンピュータ読み取り可能な記録媒体に記録される。

【０００７】

【発明の実施の形態】

本発明の実施形態では、自然言語で記述された一つ以上の文の集まりであり、且つその一つ以上の文の集まりが分類される対象である場合、それを文書と言う。具体的な例をあげれば、ＩＰＣ分類等により分類される公開特許公報や、政治・経済・文化・科学技術等の特定分野に分類される新聞記事も文書であるし、それらから請求項や特定の一文を取り出したものであっても、請求項という分類に含まれる文であるか、用途等により分類可能な特定の一文であれば文書とみなす。

以下、図面により本発明の実施の形態を詳細に説明する。

図１は本発明の第１の実施形態を示す文書分類装置の構成ブロック図である。

図１に示したように、本実施形態の文書分類装置は、文書データ群を入力する文書入力部（文書入力手段）１、それぞれの文書データを所定の基準に基づいて一つまたは複数の分割文書データに分割する文書分割部（文書分割手段）２、上記文書データと分割文書データとを対応付けるマップを生成する文書－分割文書対応マップ生成部（文書－分割文書対応マップ生成手段）３、分割文書データつまり分割された文書を分類する分割文書分類部（分割文書分類手段）４、分割文書分類結果情報を生成する分割文書分類結果生成部（分割文書分類結果生成手段）５、上記文書－分割文書対応マップと上記分割文書分類結果情報とを用いて上記文書データの分類結果情報を生成する文書分類結果生成部（文書分類結果生成手段）６などを備えている。なお、上記文書分割部２、文書－分割文書対応マップ生成部３、分割文書分類部４、分割文書分類結果生成部５、文書分類結果生成部６は共有または独自のプログラム記憶用メモリおよびプログラムに従って動作するＣＰＵを有している。

【０００８】

以下、図１などに従って、第１の実施形態の文書分類装置、文書分類方法を詳細に説明する。

まず、文書入力部１により、文書群が入力される。上記文書入力部１はキーボード、ＯＣＲ装置、着脱型記録媒体、またはネットワーク通信手段を備え、それ

らの何れか１つを介して文書データ群を入力するのである。そして、文書分割部２が上記文書データ群を取得し、それぞれの文書データを所定の基準に基づいて分割し、一つの文書データから一つまたは複数の分割文書データを生成する。なお、文書データを分割する方法としては、文書の構造情報や文書を構成する要素情報を用いたり、利用者が指定する方法などを用いるが、ここでは、その方法は問わないこととする。

図２に、この文書分類装置／文書分類方法で行われる、文書データから複数の分割文書データを生成する一例を示す。この例に示した文書１には複数のニューストピックが記述されており、１日分のトピックが文書単位となっている。図示したように、この文書ではそれぞれのニューストピックが二つの改行コードにより分離されているので、この規則を用いて一つの文書である文書１を分割し、一つが一つのトピックにより形成される分割文書1 －1 ～1 －7 の７つの分割文書データを生成する。なお、分割前の文書１も分割文書データとして含めることもできるが、ここでは含めないことにする。

文書が分割されると、文書－分割文書対応マップ生成部３が分割前の文書データとその文書データから生成された分割文書データとを対応付けるマップを生成する。例えば、個々の文書データを一意に示す識別子と個々の分割文書データを一意に示す識別子とから構成されるマップ、あるいは文書データ毎に分割文書データを一意に示す識別子から成るマップを生成するのである。なお、文書データと分割文書データを対応付ける方法についてはここでは問わないこととする。

【０００９】

図３に、文書－分割文書対応マップを生成する一例を示す。図３において、文書１～文書３は文書データを示し、分割文書１～分割文書12は分割文書データを示している。図示のように、それぞれの文書データおよび分割文書データにそれぞれを一意に識別することができる識別番号（識別子）を付与し、上記文書データの識別番号と分割文書データの識別番号とを図３の左下に示したテーブル形式で対応付けている。なお、任意の複数の分割文書データが文書分類にて用いられる基準において同一とみなすことができる場合は、それらの識別番号を同一にしてもよい。

続いて、分割文書分類部４が上記分割文書を対象に文書分類を行う。個々の分割文書に対して、例えば、言語処理を施し、文書中に含まれているそれぞれの単語の出現頻度を計数し、それに基づいてそれぞれの文書の特徴を計量的に表す特徴ベクトルを求め、それらの特徴ベクトルに対してカイ自乗法、判別分析手法、またはクラスタ分析手法などを適用することにより文書分類を行う。

次に、分割文書分類結果生成部５が上記の分割文書分類の結果に基づいた分割文書分類結果情報を生成する（図４参照）。ここで、分割文書分類結果情報とは、例えば、各分割文書データの所属カテゴリに関する情報（例えば、図４に示した「分割文書データを３つのカテゴリに分類した結果」という表中の「分類カテゴリ」および「所属カテゴリの代表値との距離」の項の情報）、生成された所属カテゴリ個々に関する情報（例えば、図４に示した「分類カテゴリに関する情報」という表中の「代表値」および「所属データ数（分割文書数）」の項の情報）、生成された所属カテゴリ間の情報（例えば表４に示した「分類カテゴリ間の距離」という表のなかの情報）などである。なお、利用者は上記のような種々の情報を分類結果分析の際の基礎データとして利用することができる。

【００１０】

図４は、12個の分割文書データをそれらの有する計量的特徴ベクトルを用いて３つのカテゴリに分類した場合の分類結果の生成例である。分割文書データの有する計量的な３次元ベクトル（ベクトルの成分数は分類対象文書群に生起するすべての単語の種類数になるが、ここでは、いくつかの単語が縮退した３次元ベクトルに線形変換している）に対して例えばクラスタ分析手法の一つであるWard法などを適用することで３つのカテゴリに分類することができる。つまり、各分割文書データは図示したように３つのカテゴリのうちのいずれか一つに属する。なお、所属カテゴリの代表値とは、所属分割文書データの特徴ベクトルの平均値（所属分割文書データの重心）である。

また、所属カテゴリの代表値との距離（類似度に対応する）は、例えば、図４の分割文書３については、分割文書データ特徴ベクトルの項における分割文書３の値と、分割文書３の分類カテゴリであるカテゴリ２の代表値（所属分割文書データの重心）の項の値により、以下の数式から求めることができる。

$$((3.00-2.66)^2+(2.00-2.00)^2+(4.00-3.66)^2)^{1/2}=0.48$$

上記の所属カテゴリの代表値との距離が小さいほど、そのカテゴリに属する平均的分割文書との類似度が高いということになる。

なお、分割文書分類結果情報としては、図４に示した以外にも、カテゴリ内分散やカテゴリ間分散、各カテゴリにおける類似度のレンジなど様々な統計量を生成することができる。

【００１１】

続いて、文書分類結果生成部６が上記文書－分割文書対応マップと上記分割文書分類結果情報とを用いて、例えば図５に示すような、上記文書データの分類結果情報を生成する。図５の例では、図示したように、各分類カテゴリ毎に、所属する分割文書データ、その類似度（所属カテゴリの代表値との距離）、分割文書データの属する分割前文書データ（所属文書）、文書占有率（分割文書データの当該カテゴリに所属する割合）、分割文書データの所属文書における相対位置（順序）、所属カテゴリ内での当該分割文書データの類似度の順位などを生成している。

なお、上記において、所属文書は文書－分割文書対応マップから、それ以外の分類結果情報は分割文書分類結果情報から得ている。文書分類結果生成部６は図５に示した情報以外にも、各カテゴリ内での分散、分割文書データの所属カテゴリ内での偏差値など様々な統計量、文書データや分割文書データの内容などを分類結果情報として利用することもできる。また、上記においては、すべての結果を分割文書データを単位とした表形式で表現しているが、分類カテゴリや文書データを単位として表現することもできる。また、分類結果情報をテキスト表現にするだけでなく、グラフィカルな表現にして、利用者が理解しやすいようにすることも可能である。

こうして、本実施形態によれば、一つの文書が分割され、分割文書が分類され、分割前文書と上記分割文書との対応が利用者に示され、上記分割文書の分類結果が利用者に示されるので、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されたりすることがなく、したがっ

て、利用者がその分類カテゴリを良く理解できる。また、分割前文書（所属文書）中の分割文書の位置なども示されるので、利用者は文書群中の読みたい部分を効率的に読むことができる。

【００１２】

図６は本発明の第２の実施形態を示す文書分類装置の構成ブロック図である。図示したように、本実施形態の文書分類装置は、図１に示した第１の実施形態の構成に加え、文書データを保存する文書保存部（文書保存手段）７、分割文書データを保存する分割文書保存部（分割文書保存手段）８、文書－分割文書対応マップ生成部３により生成された文書－分割文書対応マップを保存する文書－分割文書対応マップ保存部（文書－分割文書対応マップ保存手段）９を備えている。なお、上記各保存部は例えば共有のハードディスクや半導体メモリなどにより構成される。

上記した構成により、本実施形態の文書保存部７は、文書データの内容や、文書の作成者、作成日、最終修正日などの文書データに付随する情報を適切な形式で保存する。また、文書データが文書内容と共にその要素から成る計量的な特徴ベクトルを持つ場合にはこれらも保存する。文書入力部１にて、個々の文書データにそれらを一意に表す識別子が付与される場合にはこの識別子も適切な形式で保存することができる。

また、分割文書保存部８は、文書分割部２により生成される分割文書データの内容を適切な形式で保存すると共に、計量的な特徴ベクトルを持つ場合にはこれらも保存する。個々の上記分割文書データにそれらを一意に表す識別子が付与される場合にはこの識別子も適切な形式で保存することができる。

また、文書－分割文書対応マップ保存部９は、文書－分割文書対応マップ生成部３により生成される文書－分割文書対応マップを適切な形式で保存する。

このように、第２の実施形態によれば、文書データ、分割文書データ、および文書－分割文書対応マップが保存されるので、分割文書データおよび文書－分割文書対応マップを再生成することなしに、同一の文書データに対して、分類数、分類手法、または分類時の諸設定などパラメータの異なる分類結果を効率的に求めることができる。また、文書データを分類し、分類結果を生成するために必要

なデータが保存されることにより、利用者は、分類作業に対して時間的な自由度を持つことができ、過去に行った文書分類の再分析を任意の時間に行うこともできる。

【００１３】

図７は本発明の第３の実施形態を示す文書分類装置の構成ブロック図である。図７に示したように、本実施形態の文書分類装置は、図６に示した第２の実施形態の構成に加え、分割文書分類結果生成部５により生成された分割文書分類結果情報を保存する分割文書分類結果保存部（分割文書分類結果保存手段）10を備えている。なお、上記分割文書分類結果保存部10は、例えば、共有のハードディスクや半導体メモリなどにより構成される。

このように、第３の実施形態によれば、文書データ、分割文書データ、文書－分割文書対応マップ、および、分割文書分類結果情報が保存されるので、第２の実施形態の効果に加え、一度分類を実行すれば、その分類結果をテキスト表現や表表現やグラフ表現など様々な形式で表現することができる。また、分割文書分類結果情報が保存されることにより、分類の実行作業および分類結果の分析作業において、利用者は、時間的な自由度を持つことができ、過去に行った文書分類結果の再分析を様々な表現形式で任意の時間に行うこともできる。

本発明の第４の実施形態では、前記各実施形態の文書分類装置、文書分類方法において、図８に示すように、文書分割部２により生成される複数の分割文書データ中に分割前の文書データである文書１を含む。これにより、本実施形態では、利用者は、分割されている文書データを分類することで得られる詳細な文書データの分類構造だけでなく、分割前の文書データ自体を分類した結果として得られるマクロな分類構造の融合した分類構造を得ることができる。

【００１４】

本発明の第５の実施形態では、前記各実施形態の文書分類装置、文書分類方法において、文書分割部２は、文書データの構造情報を基に文書データを分割する。図９に、分類対象文書データがＨＴＭＬ形式で記述された文書の例を示す。分割を行う前に、図９に示したようなＨＴＭＬ形式の文書データから構造情報を抽出し、それらの構造を用いて文書の適切な分割規則を設定することにより文書デ

ータから分割文書データを生成する。つまり、この例では、文書データ中のタグ＜L1＞に着目し、「タグ＜L1＞を持つテキストを一つの分割文書データとする」という文言を分割文書データを生成する規則とする。この規則を文書データに適用することにより図９に示したような７つの分割文書が生成される。

上記のように、文書が、ＨＴＭＬ、ＸＭＬ、ＳＧＭＬなど特定の構造化文書の形式を有していない場合でも、文字の大きさ、文字の装飾、文字の色、およびフォントなどに関する情報から分割規則を生成し、分割文書を生成することもできる。また、文書データがイメージであってＯＣＲ装置などにより入力される場合には、もとのイメージのレイアウト情報などを利用することにより分割規則を生成し、分割文書を生成することもできる。

なお、文書データの全てを何れかの分割文書データにする必要はない。例えば、図９に示した例では、文字列「ニューストピック（98/09/25）」は分割文書には採用しない。

このように、第５の実施形態では、文書データから構造情報を抽出し、文書分割を行う前に構造情報を用いて文書の適切な分割規則を設定することにより、異なった話題の分割などを適切に行うことができ、したがって、文書データの詳細な分類構造がわかる文書分類を適切に行うことができる。

【００１５】

本発明の第６の実施形態では、前記第１乃至第４の実施形態の文書分類装置、文書分類方法において、図10に示すように、文書データに含まれる単語など要素を抽出する文書要素解析部（文書要素抽出手段）11、上記文書要素解析部11により抽出された要素に付随する品詞など要素付随情報を抽出する要素付随情報抽出部（要素付随情報抽出手段）12を備え（図10は図７に示した第３の実施形態に文書要素抽出部11、要素付随情報抽出12を加えた例で示している）、文書分割部２が、上記文書要素解析部11により抽出された要素、または上記要素と上記要素付随情報抽出部12により抽出された要素付随情報とを用いて上記文書データを分割する。図11に示すように、文書分割を行う前に、自然言語処理手段である文書要素解析部11が文書データから単語などそれらの要素を抽出し、要素付随情報抽出部12が品詞など要素付随情報を抽出して文書の適切な分割規則を設定するのであ

る。なお、上記文書要素解析部11および要素付随情報解析部12は新たに設けるのではなく、分割文書分類部４内の同様の手段を用いることが可能である。

本実施形態では、例えば、図11に示したように、文書データが特定の構造情報を持たない複数のニューストピックの集まりであり、各トピックが、単語「トピック」＋「数字」＋「改行コード」という文字列の後に記述されている場合で説明すると、上記のような構造が文書要素解析部11および要素付随情報抽出部12の抽出結果から認識され、文章の終端を考慮して、「トピック＋数字＋改行コードという文字列を先頭とし、上記文字列または文書終端記号を終端として囲まれる文字列を一つの分割文書データとする」という分割規則が生成されることになる。

さらに詳しく説明すると、抽出された単語とその品詞情報などから、まず、名詞と改行コードのみを抽出し、次に、文字列「トピック＋数字＋改行コード」および文書終端記号を検出し、文書内でのそれらの位置を記憶する。そして、文書データに対して前記分割規則を適用し、図11に示したような分割文書データを生成する。

なお、文書データのすべてをいずれかの分割文書データにする必要はなく、例えば、図11に示した例では、文字列「ニューストピック（98/09/25）」は分割文書には採用しない。また、上記の例では、文書データから要素およびその付随情報を抽出して分割規則を設定する場合で説明したが、要素のみを抽出してその要素情報から分割規則を設定することも可能である。

こうして、第６の実施形態によれば、文書データからそれらの要素情報などを抽出し、抽出した要素情報などを用いて文書の分割規則を設定することにより、第５の実施形態と同様に、文書データの詳細な分類構造がわかる文書分類を適切に行うことができる。

【００１６】

本発明の第７の実施形態では、前期第１乃至第４の実施形態の文書分類装置、文書分類方法において、利用者により指示された指定範囲に従って文書分割部２が文書データを分割する。図12に示すような文書データに対して利用者がそれぞれの分割文書の範囲を指定すると、指定に従って文書分割部２が文書分割を行う

本実施形態では、文書分割時、文書分割部２がまず、画面上に、その初期状態として左右の指示ポイントおよび領域指定ラインから成る領域指定オブジェクトを文書の最上部に表示する。この状態で、利用者は、マウスなどポインティングデバイスを用いて、左右どちらかの指示ポイントをドラッグし、それを上下に移動させることにより、それぞれの分割文書の領域を選択することができる。また、この指定時、文書分割部２は、領域選択処理を行っていることを示すため、指示ポインタを黒色から白色に、領域指定ラインを実線から破線に変化させる。選択領域を決定するには、所望の位置で指示ポイントのドラッグを止めればよい。

次に、利用者は選択した領域を分割文書とするかしないか決定する。分割領域としない場合には、それを明示的に表示するために、文書分割部２は選択領域を図示のように網掛け表示にさせる。

こうして、本実施形態によれば、利用者は文書データからそれぞれの分割文書データを所望通りに選択することができるので、文書データの詳細な分類構造がわかり、且つ利用者の意図に合った文書分類を行うことができる。

【００１７】

本発明の第８の実施形態では、前期第１乃至第４の実施形態の文書分類装置、文書分類方法において、文書データ中の文字数、文数、または文字数と文数の両方を基に文書データを分割する。例えば、図13に示す文書データをほぼ200 文字を単位として分割を行う。ここで、ほぼ200 文字を単位とするのは、正確な200 文字単位としてもその終端が句点である保証がないことから、200 文字目の前後の最も近い句点をそれぞれの分割文書の終端とするからである。こうして、図13に示したような分割文書が生成される。

同様に、所定の文数を単位とした文書分割を行うこともできるし、文字数と文数の両方を基にした文書分割を行うこともできる。

このように、第８の実施形態によれば、文字数、文数、または文字数と文数の両方を基に文書データを分割することにより、話題の異なった内容などが異なった分割文書として分割され、分類される可能性が高くなるので、文書データの詳細な分類構造がわかる文書分類を行うことができる。

本発明の第９の実施形態では、前記各実施形態の文書分類装置、文書分類方法において、文書分類結果生成部６が分類結果情報として、文書データを示す情報および上記文書データに付随する代表的情報のみを提示する。例えば図14に示すように、先頭に分類カテゴリ名を表示し、その横にそのカテゴリを代表するキーワードを表示し、カテゴリ名の下には文書データを示す情報として当該カテゴリに属する分割文書データを含んでいる文書データの、例えば、文書データ名（文書名）を表示する。また、各文書データ名の左側には文書アイコンを表示させ、この文書アイコンが指示されたとき、文書データの内容を表示させる。また、各文書データ名の配置順は、カテゴリ代表値との類似度が高い分割文書データの文書データ名を先（左側）にする。また、同じ文書データから生成された複数の分割文書データが同一の分類カテゴリに属している場合には、類似度の最も高い分割文書データに対応する文書データ名のみを表示する。なお、上記キーワードとは出現頻度の多い単語である。

このように、第９の実施形態によれば、文書分類結果が文書データを示す情報と文書データに付随する代表的情報のみが表示されるので、利用者は文書データの詳細な分類構造の概要を容易に把握することができる。

【００１８】

本発明の第10の実施形態では、第９の実施形態の文書分類結果提示に加えて、分割文書データを示す情報および上記分割文書データに付随する情報を提示する。例えば、図15に示すように、先頭に分類カテゴリ名を表示し、その横にそのカテゴリを代表するキーワードを表示し、カテゴリ名の下には文書データを示す情報として当該カテゴリに属する分割文書データを含んでいる文書データの例えば文書データ名（文書名）を表示する。

また、各文書データ名の左側には文書アイコンを表示させ、この文書アイコンが指示されたとき、文書データの内容を表示させる。また、文書データ名の右側には分割文書アイコンを表示させる。なお、分割文書アイコン中には当該文書データにおける分割文書データの位置と当該文書データ中の分割文書数を表示させる。さらに、上記分割文書アイコンを指示することで文書データ中の当該分割文書データを表示させることができる。

また、各文書データ名の配置順はカテゴリ代表値との類似度が高い分割文書データの文書データ名を先にする。また、同じ文書データから生成された複数の分割文書データが同一の分類カテゴリに属している場合には類似度の順位がわかるようにその順位を表示させる。

このように、第10の実施形態によれば、文書分類結果が文書データを示す情報と文書データに付随する代表的情報、および分割文書データを示す情報と分割文書データに付随する代表的情報のみが表示されるので、利用者は文書データの詳細な分類構造の概要と共にどの分割文書が起因して当該カテゴリに分類されたかというようなことも容易にわかる。

以上、本発明の文書分類装置および文書分類方法を説明したが、この文書分類方法を実現するプログラムを着脱可能であると共にコンピュータ読み取り可能な記録媒体に記録し、上記記録媒体を移した先の情報処理装置内で本発明によった文書分類を行うこともできる。

【００１９】

【発明の効果】

以上説明したように、本発明によれば、請求項１および請求項11記載の発明では、入力された文書データ群の各文書が分割され、一つの文書データから一つまたは複数の分割文書データが生成され、上記文書データと上記分割文書データとの対応を示す文書－分割文書対応マップが生成され、上記分割文書データが分類され、分割文書分類結果情報が生成され、上記文書－分割文書対応マップと上記分割文書分類結果情報とを用いて上記文書データの分類結果情報が生成されるので、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されたりすることがなく、したがって、利用者がその分類カテゴリを良く理解できる。また、分割前文書（所属文書）中の分割文書の位置なども示されるので、利用者は文書群中の読みたい部分を効率的に読むことができる。

また、請求項２記載の発明では、請求項１記載の発明において、文書データ、分割文書データ、および文書－分割文書対応マップが保存されるので、分割文書

データおよび文書－分割文書対応マップを再生成することなしに、同一の文書データに対して、分類数、分類手法、または分類時の諸設定などパラメータの異なる分類結果を効率的に求めることができる。また、文書データを分類し、分類結果を生成するために必要なデータが保存されることにより、利用者が分類作業に対して時間的な自由度を持つことができるし、過去に行った文書分類の再分析を任意の時間に行うこともできる。

また、請求項３記載の発明では、請求項２記載の発明において、さらに、分割文書分類結果情報が保存されるので、請求項２記載の発明の効果に加え、一度分類を実行すれば、その分類結果をテキスト表現や表表現やグラフ表現など様々な形式で表現することができる。また、分割文書分類結果情報が保存されることにより、分類の実行作業および分類結果の分析作業において、利用者が時間的な自由度を持つことができるし、過去に行った文書分類結果の再分析を様々な表現形式で任意の時間に行うこともできる。

また、請求項４記載の発明では、請求項１、請求項２または請求項３記載の発明において、複数の分割文書データのなかに分割前の文書データそのものが含まれるので、利用者は、分割されている文書データを分類することで得られる詳細な文書データの分類構造だけでなく、分割前の文書データ自体を分類した結果として得られる概略的でマクロな分類構造の融合した分類構造を得ることができる。

また、請求項５記載の発明では、請求項１乃至請求項４記載の発明において、文書データの構造情報を基にして文書データが分割されるので、異なった話題の分割等を適切に行うことができ、したがって、文書データの詳細な分類構造がわかる文書分類を適切に行うことができる。

【００２０】

また、請求項６記載の発明では、請求項１乃至請求項４記載の発明において、分割対象の文書データから抽出された要素、または上記要素と上記要素から抽出された要素付随情報とを用いて上記文書データが分割されるので、請求項５記載の発明と同様に、文書データの詳細な分類構造がわかる文書分類を適切に行うことができる。

また、請求項７記載の発明では、請求項１乃至請求項４記載の発明において、利用者により指示された指定範囲に従って文書データが分割されるので、利用者の意図に合い、且つ文書データの詳細な分類構造がわかる文書分類を行うことができる。

また、請求項８記載の発明では、請求項１乃至請求項４記載の発明において、文書データ中の文字数、文数、または文字数と文数の両方を基に文書データが分割されるので、話題の異なった内容などが異なった文書として分類される可能性が高くなり、したがって、この発明でも文書データの詳細な分類構造がわかる文書分類を行うことができる。

また、請求項９記載の発明では、請求項１乃至請求項８記載の発明において、分類結果情報として、文書データを示す情報および上記文書データに付随する代表的情報が抽出・提示されるので、利用者は文書データの詳細な分類構造の概要や全体的な構造を容易に把握することができる。

また、請求項10記載の発明では、請求項９記載の発明において、分類結果情報として、さらに、分割文書データを示す情報および上記分割文書データに付随する代表的情報が抽出・提示されるので、利用者は文書データの詳細な分類構造の概要や全体的な構造と共にどの分割文書が起因して当該カテゴリに分類されたかというようなことも容易にわかる。

また、請求項12記載の発明では、請求項11記載の発明によったプログラムがコンピュータ読み取り可能な記録媒体に記録されるので、上記記録媒体を移した先の情報処理装置内で本発明によった文書分類を行うこともできる。

【図面の簡単な説明】

【図１】

本発明の第１の実施形態を示す文書分類装置の構成ブロック図である。

【図２】

本発明の第１の実施形態を示す文書分類装置および文書分類方法の説明図である。

【図３】

本発明の第１の実施形態を示す文書分類装置および文書分類方法の他の説明図

である。

【図４】

本発明の第１の実施形態を示す文書分類装置および文書分類方法の他の説明図である。

【図５】

本発明の第１の実施形態を示す文書分類装置および文書分類方法の他の説明図である。

【図６】

本発明の第２の実施形態を示す文書分類装置の構成ブロック図である。

【図７】

本発明の第３の実施形態を示す文書分類装置の構成ブロック図である。

【図８】

本発明の第４の実施形態を示す文書分類装置および文書分類方法の説明図である。

【図９】

本発明の第５の実施形態を示す文書分類装置および文書分類方法の説明図である。

【図10】

本発明の第６の実施形態を示す文書分類装置の構成ブロック図である。

【図11】

本発明の第６の実施形態を示す文書分類装置および文書分類方法の説明図である。

【図12】

本発明の第７の実施形態を示す文書分類装置および文書分類方法の説明図である。

【図13】

本発明の第８の実施形態を示す文書分類装置および文書分類方法の説明図である。

【図14】

本発明の第９の実施形態を示す文書分類装置および文書分類方法の説明図である。

【図15】

本発明の第10の実施形態を示す文書分類装置および文書分類方法の説明図である。

【符号の説明】

1　文書入力部

2　文書分割部

3　文書－分割文書対応マップ生成部

4　分割文書分類部

5　分割文書分類結果生成部

6　文書分類結果生成部

7　文書保存部

8　分割文書保存部

9　文書－分割文書対応マップ保存部

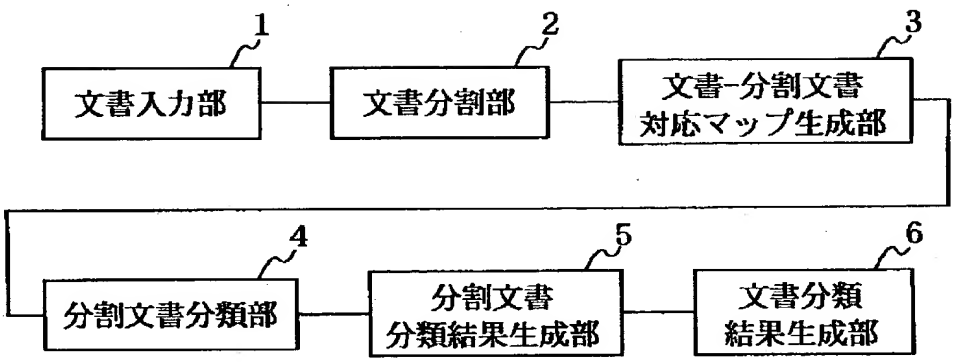
10　分割文書分類結果保存部

11　文書要素解析部

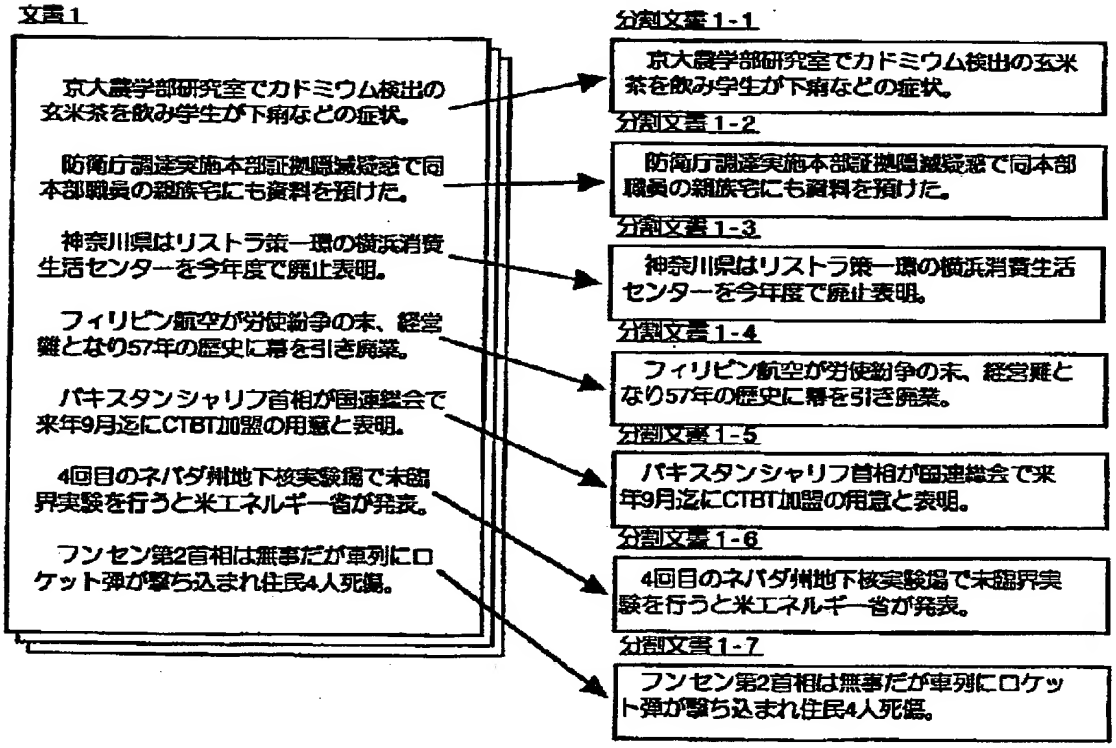
12　要素付随情報抽出部

【書類名】　図面

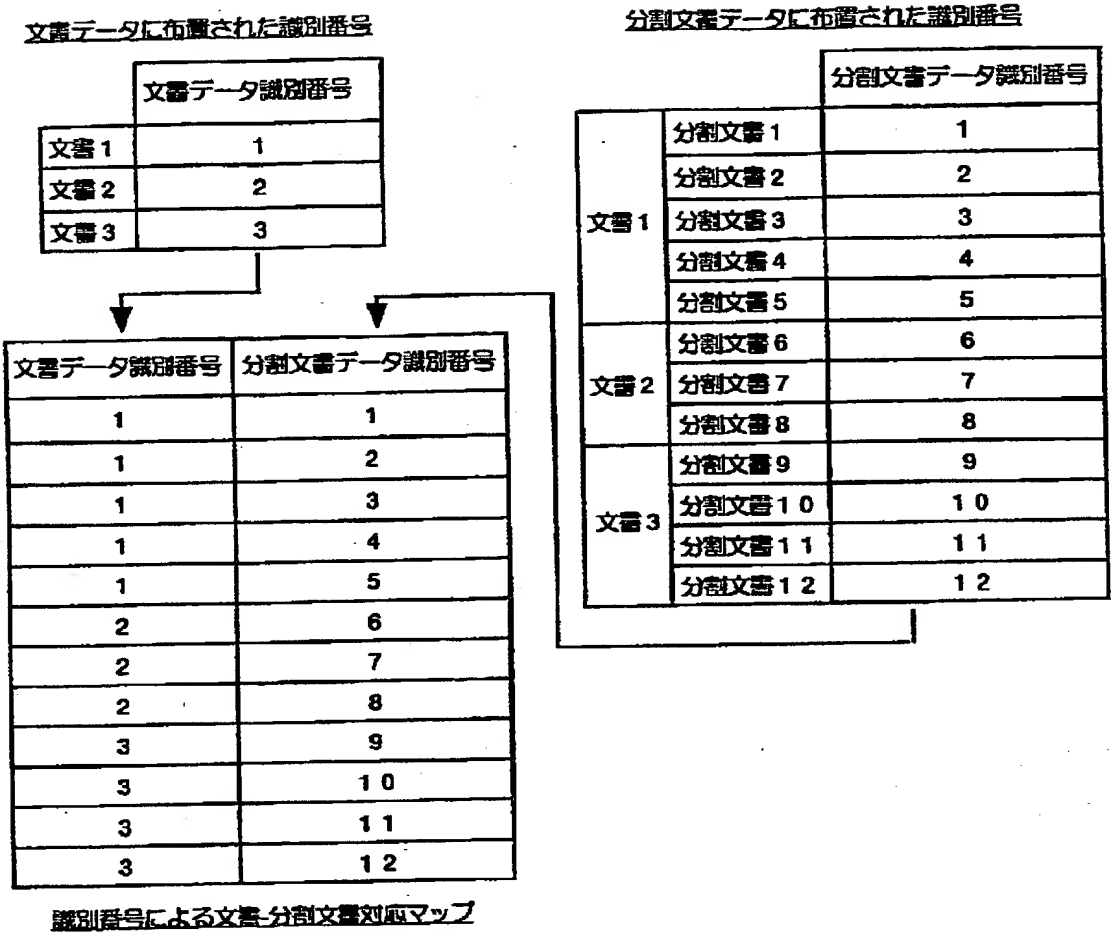
【図１】


1
文書入力部
2
文書分割部
3
文書−分割文書
対応マップ生成部
4
分割文書分類部
5
分割文書
分類結果生成部
6
文書分類
結果生成部

【図２】


文書1
京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。
分割文書1-1
京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
分割文書1-2
防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
分割文書1-3
神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
分割文書1-4
フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
分割文書1-5
パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
分割文書1-6
4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
分割文書1-7
フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。

【図３】


文書データに布置された識別番号
文書データ識別番号
文書１ 1
文書２ 2
文書３ 3
分割文書データに布置された識別番号
分割文書データ識別番号
文書１ 分割文書１ 1
分割文書２ 2
分割文書３ 3
分割文書４ 4
分割文書５ 5
文書２ 分割文書６ 6
分割文書７ 7
分割文書８ 8
文書３ 分割文書９ 9
分割文書１０ １０
分割文書１１ １１
分割文書１２ １２
文書データ識別番号 分割文書データ識別番号
1 1
1 2
1 3
1 4
1 5
2 6
2 7
2 8
3 9
3 １０
3 １１
3 １２
識別番号による文書-分割文書対応マップ

【図４】

分割文書データの特徴ベクトル表現

	分割文書データ識別番号	分割文書データ特徴ベクトル
分割文書１	１	（１、１、１）
分割文書２	２	（５、５、５）
分割文書３	３	（３、２、４）
分割文書４	４	（３、２、３）
分割文書５	５	（５、４、６）
分割文書６	６	（１、２、１）
分割文書７	７	（１、０、１）
分割文書８	８	（５、４、５）
分割文書９	９	（２、２、４）
分割文書１０	１０	（２、１、１）
分割文書１１	１１	（４、４、６）
分割文書１２	１２	（５、５、６）

文書分類（クラスタ分析手法を適用）

分割文書データを３つのカテゴリに分類した結果

	分割文書データ識別番号	分類カテゴリ	所属カテゴリの代表値との距離
分割文書１	１	カテゴリ１	０．２５
分割文書２	２	カテゴリ３	０．８７
分割文書３	３	カテゴリ２	０．４８
分割文書４	４	カテゴリ２	０．７４
分割文書５	５	カテゴリ３	０．５４
分割文書６	６	カテゴリ１	１．０３
分割文書７	７	カテゴリ１	１．０３
分割文書８	８	カテゴリ３	０．７０
分割文書９	９	カテゴリ２	０．７４
分割文書１０	１０	カテゴリ１	０．７５
分割文書１１	１１	カテゴリ３	０．９４
分割文書１２	１２	カテゴリ３	０．８３

分類カテゴリに関する情報

	代表値（所属分割文書データの重心）	所属データ数
カテゴリ１	（１．２５、１．００、１．００）	４
カテゴリ２	（２．６６、２．００、３．６６）	３
カテゴリ３	（４．８０、４．４０、５．６０）	５

分類カテゴリ間の距離

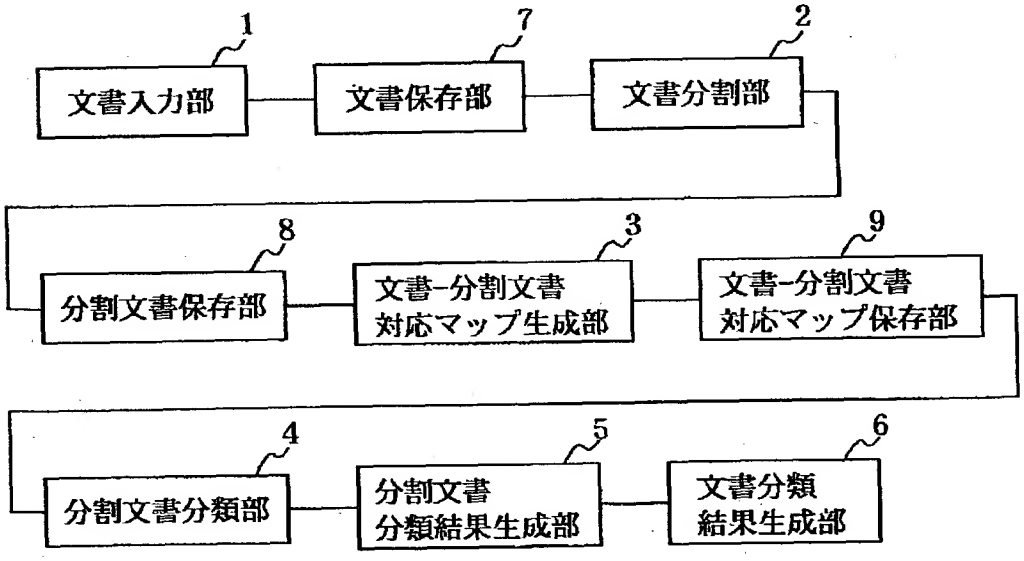
	カテゴリ２	カテゴリ３
カテゴリ１	３．１７	６．６８
カテゴリ２		３．６９

【図５】

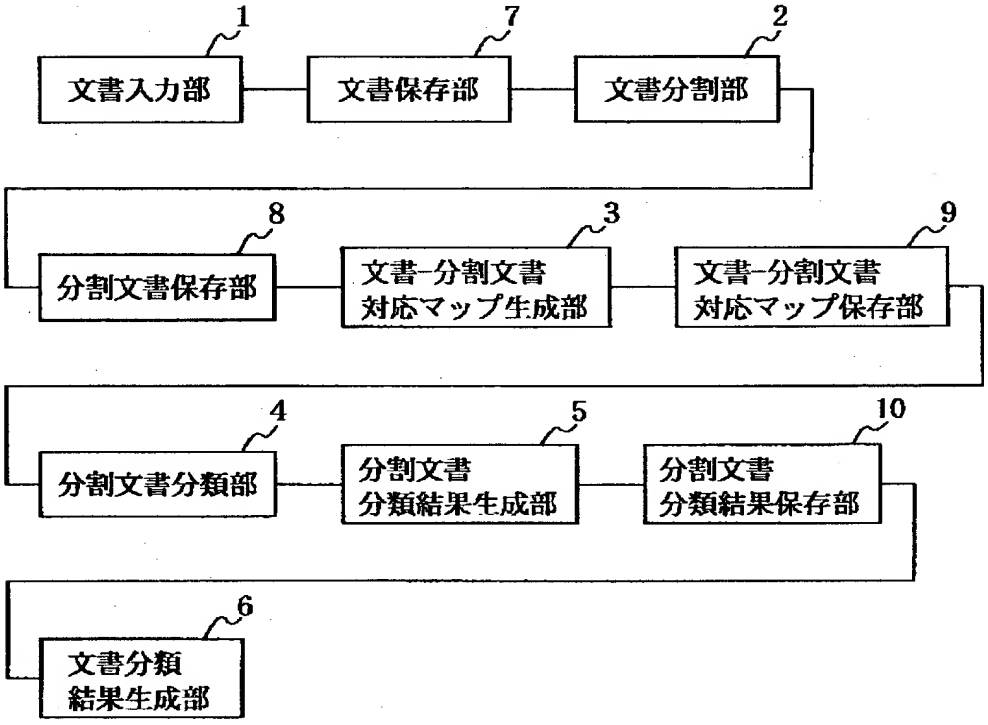
文書分類結果

分類カテゴリ	分割文書	類似度	所属文書	文書占有率	相対位置	類似順位
カテゴリ１	分割文書1	0.25	文書１	１／５	１／５	1
カテゴリ１	分割文書６	1.03	文書２	２／３	１／３	3
カテゴリ１	分割文書７	1.03	文書２	２／３	２／３	3
カテゴリ１	分割文書１０	0.75	文書３	１／４	２／４	2
カテゴリ２	分割文書３	0.48	文書１	２／５	３／５	1
カテゴリ２	分割文書４	0.74	文書１	２／５	４／５	2
カテゴリ２	分割文書９	0.74	文書３	１／４	１／４	2
カテゴリ３	分割文書２	0.87	文書１	２／５	２／５	4
カテゴリ３	分割文書５	0.54	文書１	２／５	５／５	1
カテゴリ３	分割文書８	0.70	文書２	１／３	３／３	2
カテゴリ３	分割文書１１	0.94	文書３	２／４	３／４	4
カテゴリ３	分割文書１２	0.83	文書３	２／４	４／４	3

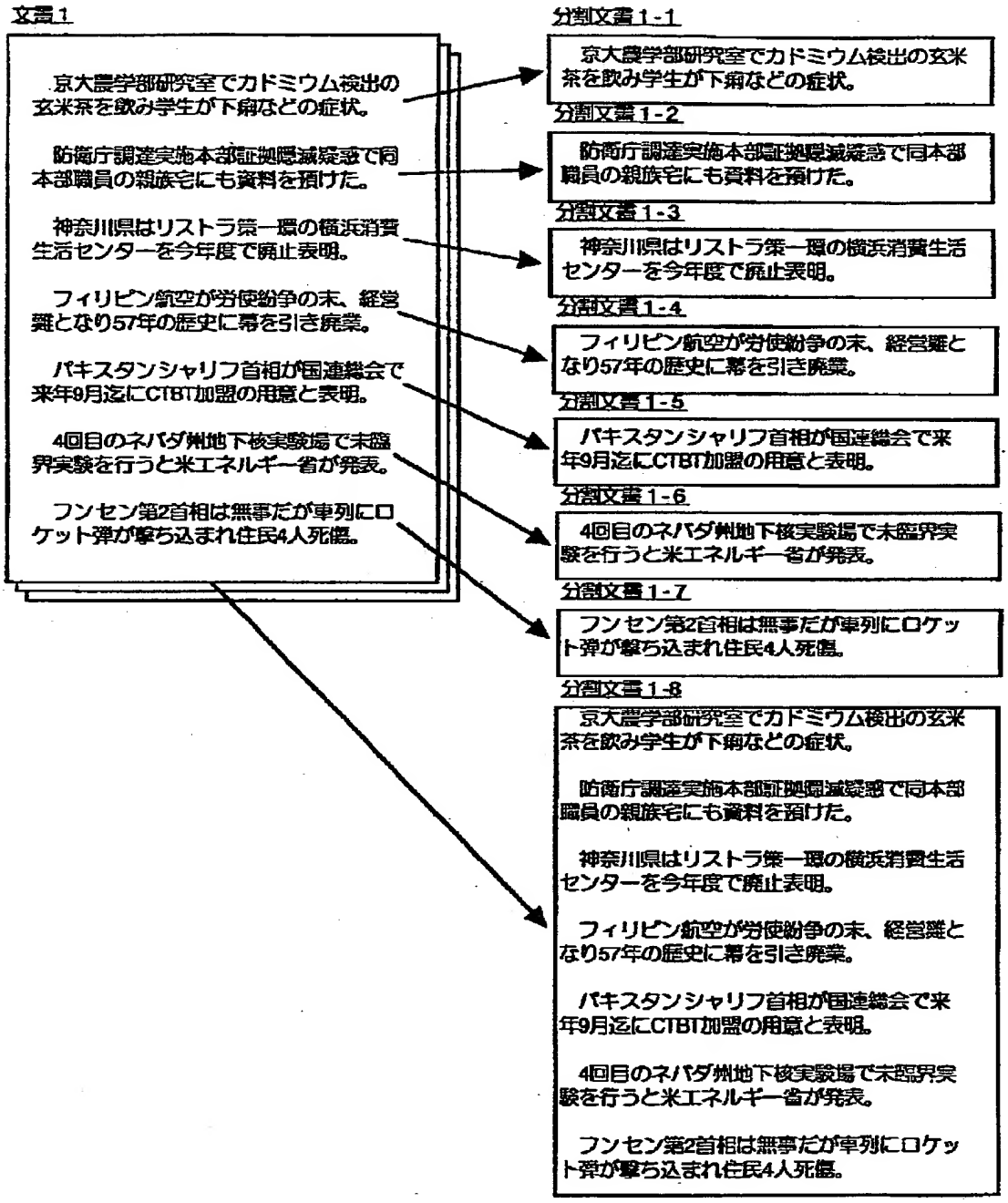
【図６】



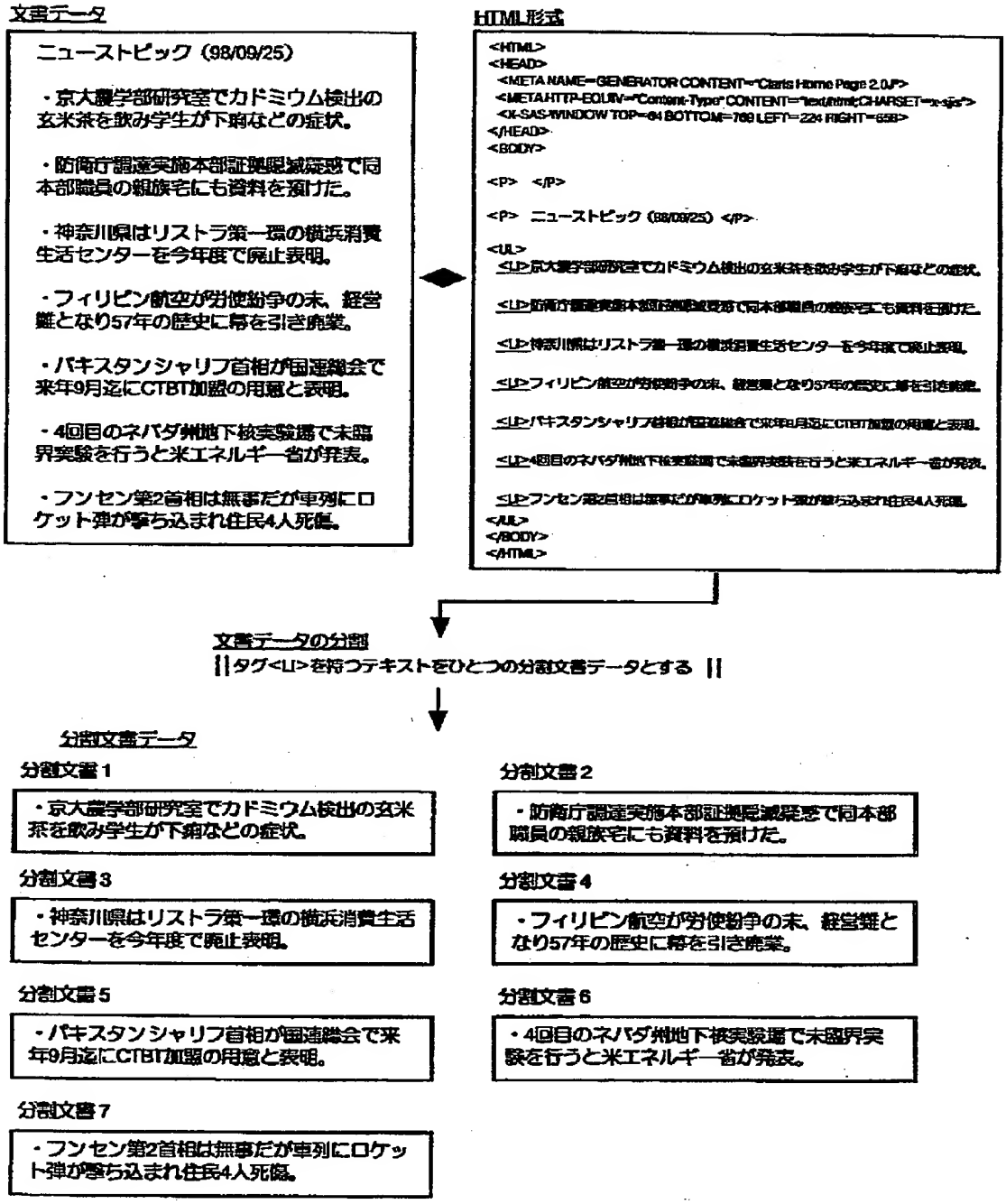
【図７】


1
文書入力部
7
文書保存部
2
文書分割部
8
分割文書保存部
3
文書−分割文書
対応マップ生成部
9
文書−分割文書
対応マップ保存部
4
分割文書分類部
5
分割文書
分類結果生成部
10
分割文書
分類結果保存部
6
文書分類
結果生成部

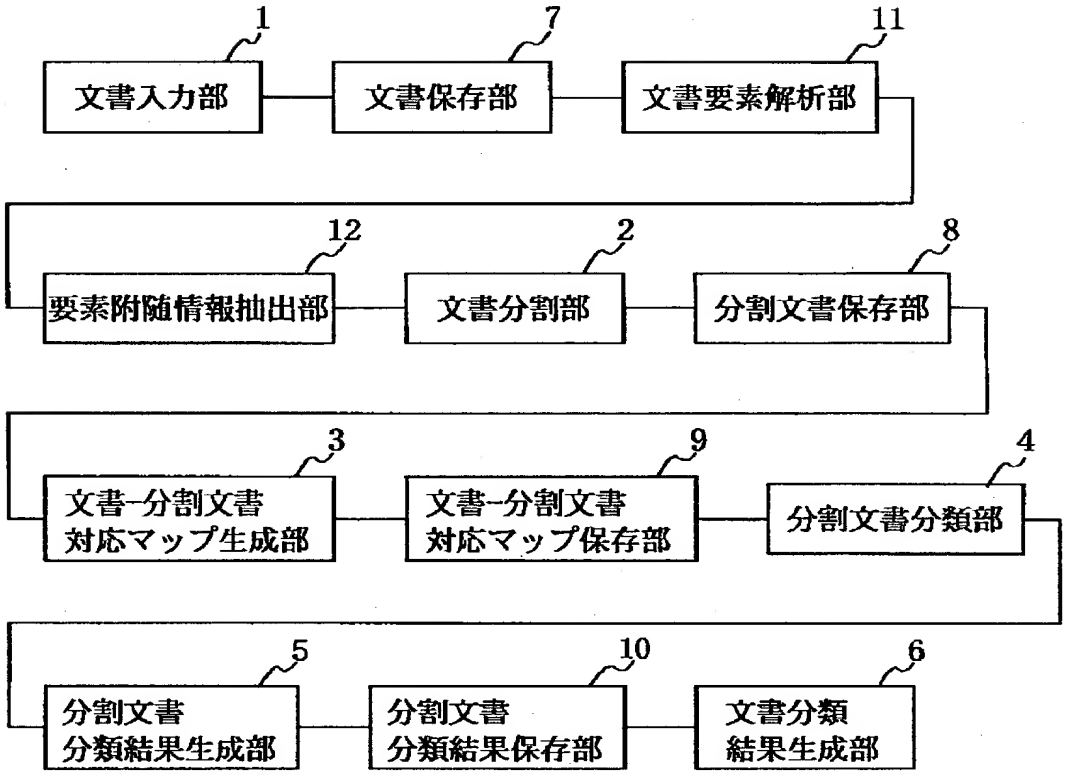
【図８】


文書１
京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。
分割文書１-１
京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
分割文書１-２
防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
分割文書１-３
神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
分割文書１-４
フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
分割文書１-５
パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
分割文書１-６
4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
分割文書１-７
フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。
分割文書１-８
京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。

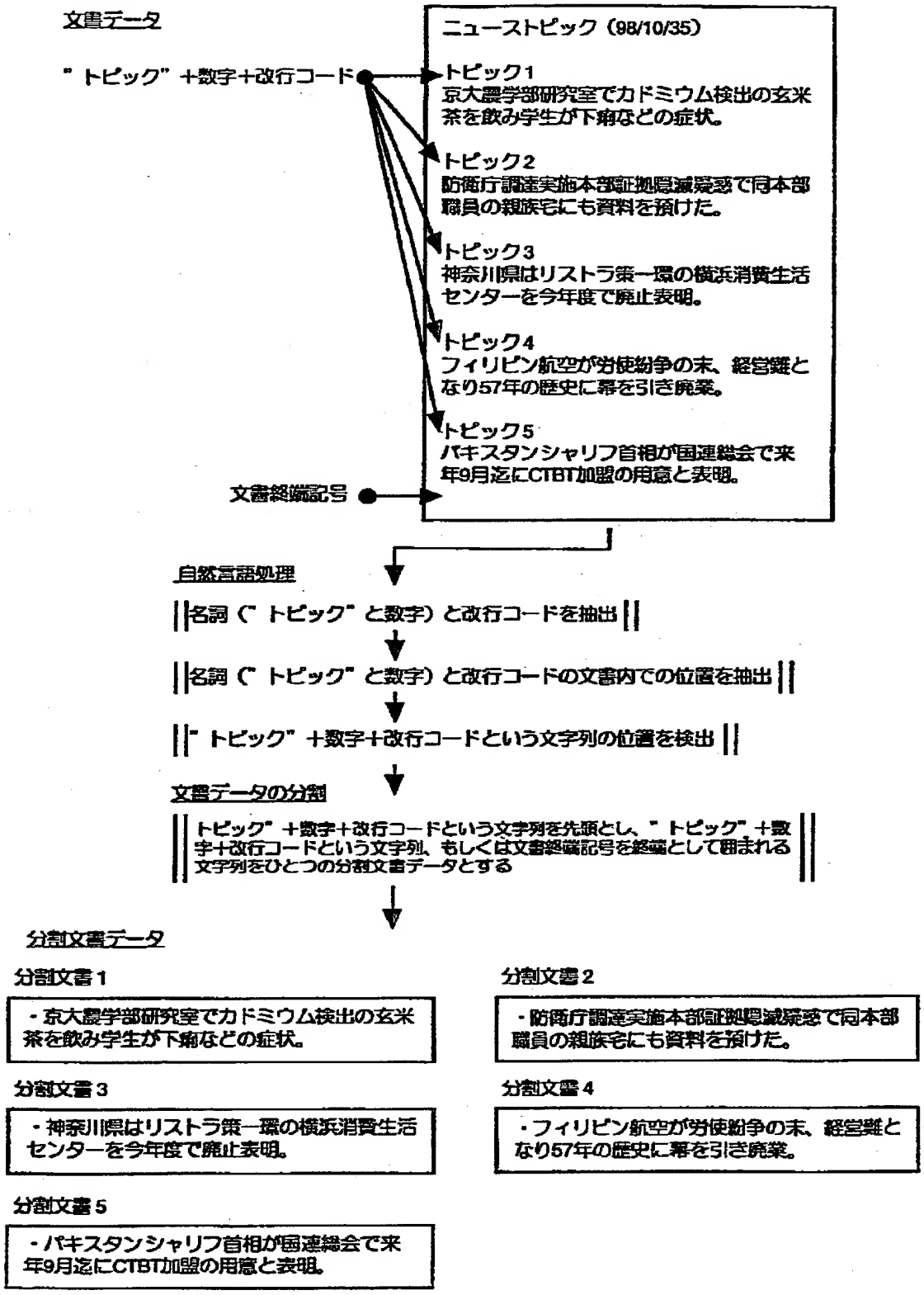
【図９】


文書データ
ニューストピック（98/09/25）
・京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
・防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
・神奈川県はリストラ第一環の横浜消費生活センターを今年度で廃止表明。
・フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
・パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
・4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
・フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。
HTML形式
<HTML>
<HEAD>
<META NAME="GENERATOR CONTENT="Claris Home Page 2.0/J">
<META HTTP-EQUIV="Content-Type" CONTENT="text/html;CHARSET="x-sjis">
<X-SAS-WINDOW TOP=64 BOTTOM=769 LEFT=224 RIGHT=658>
</HEAD>
<BODY>
<P> </P>
<P> ニューストピック（98/09/25）</P>
<UL>
<LI>京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
<LI>防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
<LI>神奈川県はリストラ第一環の横浜消費生活センターを今年度で廃止表明。
<LI>フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
<LI>パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
<LI>4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
<LI>フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。
</UL>
</BODY>
</HTML>
文書データの分割
||タグ<LI>を持つテキストをひとつの分割文書データとする ||
分割文書データ
分割文書１
・京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
分割文書２
・防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
分割文書３
・神奈川県はリストラ第一環の横浜消費生活センターを今年度で廃止表明。
分割文書４
・フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
分割文書５
・パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
分割文書６
・4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
分割文書７
・フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。

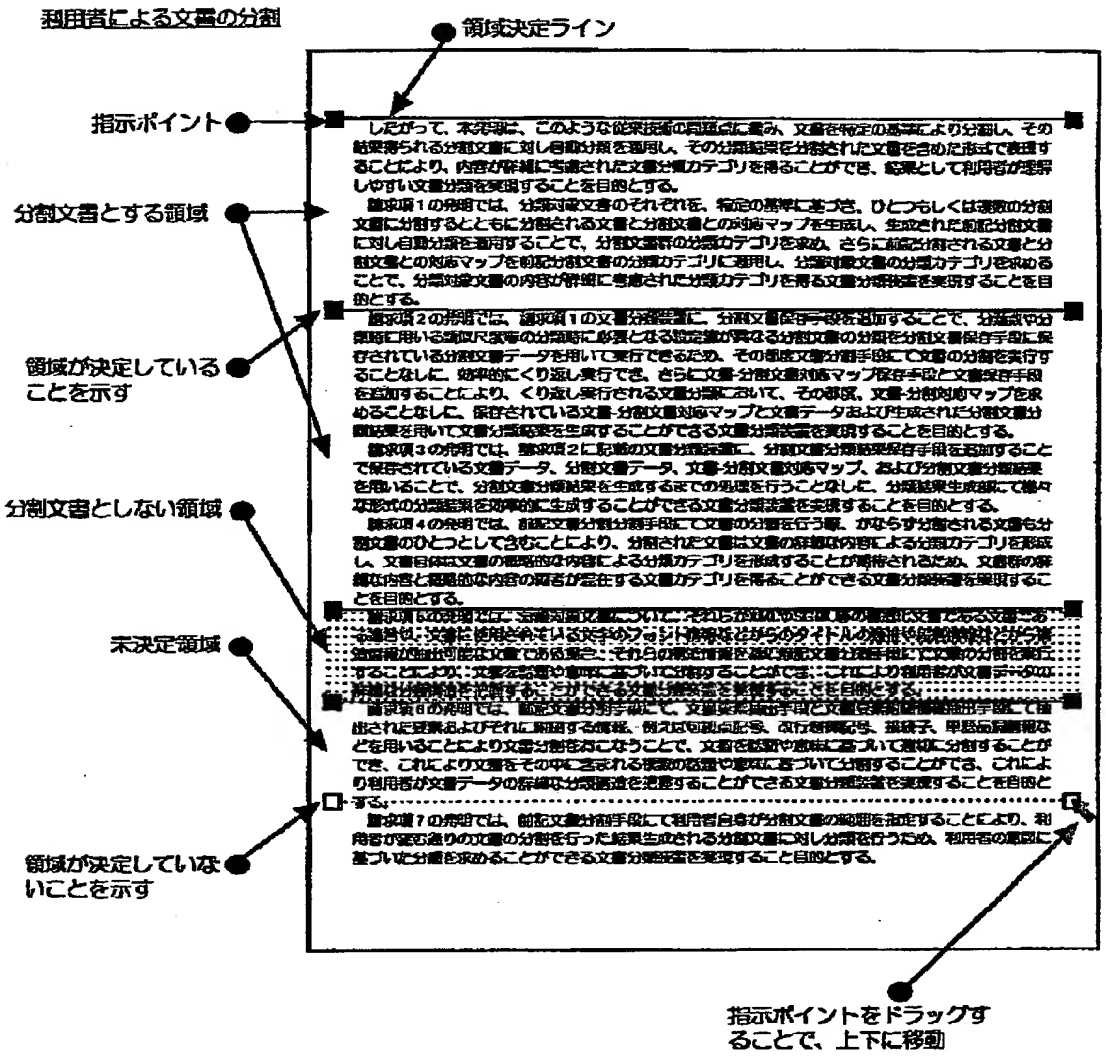
【図１０】


1
文書入力部
7
文書保存部
11
文書要素解析部
12
要素附随情報抽出部
2
文書分割部
8
分割文書保存部
3
文書−分割文書
対応マップ生成部
9
文書−分割文書
対応マップ保存部
4
分割文書分類部
5
分割文書
分類結果生成部
10
分割文書
分類結果保存部
6
文書分類
結果生成部

【図１１】


文書データ
" トピック" +数字+改行コード
ニューストピック（98/10/35）
トピック1
京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
トピック2
防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
トピック3
神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
トピック4
フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
トピック5
パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
文書終端記号
自然言語処理
名詞（" トピック" と数字）と改行コードを抽出
名詞（" トピック" と数字）と改行コードの文書内での位置を抽出
" トピック" +数字+改行コードという文字列の位置を検出
文書データの分割
トピック" +数字+改行コードという文字列を先頭とし、" トピック" +数字+改行コードという文字列、もしくは文書終端記号を終端として囲まれる文字列をひとつの分割文書データとする
分割文書データ
分割文書1
・京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
分割文書2
・防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
分割文書3
・神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
分割文書4
・フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
分割文書5
・パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。

【図１２】


利用者による文書の分割
領域決定ライン
指示ポイント
分割文書とする領域
領域が決定している
ことを示す
分割文書としない領域
未決定領域
領域が決定していな
いことを示す
したがって、本発明は、このような従来技術の問題点に鑑み、文書を特定の基準により分割し、その結果得られる分割文書に対し自動分類を適用し、その分類結果を分類された文書を含めた形式で表現することにより、内容が詳細に考慮された文書分類のテゴリを得ることができ、結果として利用者が理解しやすい文書分類を実現することを目的とする。
請求項１の発明では、分類対象文書のそれぞれを、特定の基準に基づき、ひとつもしくは複数の分割文書に分割するとともに分割される文書と分割文書との対応マップを生成し、生成された前記分割文書に対し自動分類を適用することで、分割文書群の分類のテゴリを求め、さらに前記分割される文書と分割文書との対応マップを前記分割文書の分類カテゴリに適用し、分類対象文書の分類カテゴリを求めることで、分類対象文書の内容が詳細に考慮された分類カテゴリを得る文書分類処理を実現することを目的とする。
請求項２の発明では、請求項１の文書分類装置に、分割文書保存手段を追加することで、分類器や分類時に用いる類似尺度等の分類時に必要となる設定値が異なる分割文書の分類を分割文書保存手段に保存されている分割文書データを用いて実行できるため、その類似文書分類手段にて文書の分類を実行することなしに、効率的にくり返し実行でき、さらに文書-分割文書対応マップ保存手段と文書保存手段を追加することにより、くり返し実行される文書分類において、その都度、文書-分割対応マップを求めることなしに、保存されている文書-分割文書対応マップと文書データおよび生成された分割文書分類結果を用いて文書分類結果を生成することができる文書分類装置を実現することを目的とする。
請求項３の発明では、請求項２に記載の文書分類装置に、分割文書分類結果保存手段を追加することで保存されている文書データ、分割文書データ、文書-分割文書対応マップ、および分割文書分類結果を用いることで、分割文書分類結果を生成するまでの処理を行うことなしに、分類結果生成部にて様々な形式の分類結果を効率的に生成することができる文書分類装置を実現することを目的とする。
請求項４の発明では、前記文書分割分類手段にて文書の分割を行う際、かならず分割される文書も分割文書のひとつとして含むことにより、分割された文書は文書の詳細な内容による分類カテゴリを形成し、文書自体は文書の概略的な内容による分類カテゴリを形成することが可能されるため、文書群の詳細な内容と概略的な内容の両者が混在する文書カテゴリを得ることができる文書分類装置を実現することを目的とする。
請求項５の発明では、文書対象文書について、それらが他の文書の書誌的な要素である文書であるか否か、文書に使用されている文字のフォント情報はどうかタイトルの有無や統計的な特徴量が何か等の抽出可能な文書であるか、それらの属性情報を前記分類文書生成手段にて文書の分割を実行することにより、文書を記憶や意味に基づいて分類することができ、これにより利用者の文書データの所望の分類結果を得ることができる文書分類装置を実現することを目的とする。
請求項６の発明では、前記文書分割手段にて、文章表現辞書手段と文章表現照合手段にて抽出された要素およびそれに関連する属性、例えば句読点記号、改行記号、括弧子、単語辞書情報などを用いることにより文書分割を行なうことで、文書を記憶や意味に基づいて適切に分割することができ、これにより文書をその中に含まれる複数の記憶や意味に基づいて分割することができ、これにより利用者が文書データの詳細な分類構造を把握することができる文書分類装置を実現することを目的とする。
請求項７の発明では、前記文書分割手段にて利用者自身が分割文書の細部を指定することにより、利用者が望む通りの文書の分割を行った結果生成される分割文書に対し分類を行うため、利用者の意図に基づいた分類を求めることができる文書分類装置を実現すること目的とする。
指示ポイントをドラッグすることで、上下に移動

【図１３】

文字数と文数による分割

文書データ

ユーザの意図を反映するような文書分類をおこなうためのひとつの方法として、前記表現空間変換関数により構成される空間における不必要な特徴次元や、悪影響を及ぼすような特徴次元に対し削除や合成をおこなったり、逆にある特徴次元を強調させるための操作をすることが考えられる。しかし、前記表現空間変換関数により生成される空間の特徴次元は、前記文書群解析部にて抽出される単語のうち意味的に似たものが複数結合したものと考えることができるため、各特徴次元の意味的な解釈は極めて複雑かつ多義的なものであるため、ユーザに各特徴次元の意味を提示することは極めて難しい。そこで、ユーザに分類に反映させたくない内容や強調したい内容をもつ文書や単語などの情報を指定させ、それらを前記表現空間辺関数により構成される空間に適切に射影し、それらと類似度の高い特徴次元や低い特徴次元を判別することで、操作をおこなう特徴次元を選択することが考えられる。ここでは、前記表現空間変換関数の特徴次元を操作する例として、ユーザが指定するある文書と類似度の高い特徴次元の削除を行う例を示す。ユーザにより指定された文書を前記文書特徴ベクトルと同じ次元数をもつベクトルで表現し、その文書ベクトルに前記表現空間変換関数を適用し文書ベクトルを前記表現空間変換関数により構成される空間へ射影する。そして、この射影された文書ベクトルと各特徴次元との類似度を算出することで、類似度の高い特徴次元を判別する。このとき、類似度を測るための尺度としては、余弦尺度、内積尺度、ユークリッド距離尺度などを用いることができる。また、判別に関しては、ある類似度以上を削除対象として採用するような閾値処理よる判別、類似度の高い順にある一定数を削除対象として採用する定数処理、もしくは判別分析なども用いることができる。このようにして、採用された特徴次元を前記表現空間変換関数から削除することで前記表現空間変換関数を修正することができる。

↓

文書データの分割

先頭から２００文字目の文字からその前後で最も近い句点までをひとつの分割文書とする。

↓

分割文書１

ユーザの意図を反映するような文書分類をおこなうためのひとつの方法として、前記表現空間変換関数により構成される空間における不必要な特徴次元や、悪影響を及ぼすような特徴次元に対し削除や合成をおこなったり、逆にある特徴次元を強調させるための操作をすることが考えられる。しかし、前記表現空間変換関数により生成される空間の特徴次元は、前記文書群解析部にて抽出される単語のうち意味的に似たものが複数結合したものと考えることができるため、各特徴次元の意味的な解釈は極めて複雑かつ多義的なものであるため、ユーザに各特徴次元の意味を提示することは極めて難しい。

分割文書２

そこで、ユーザに分類に反映させたくない内容や強調したい内容をもつ文書や単語などの情報を指定させ、それらを前記表現空間辺関数により構成される空間に適切に射影し、それらと類似度の高い特徴次元や低い特徴次元を判別することで、操作をおこなう特徴次元を選択することが考えられる。ここでは、前記表現空間変換関数の特徴次元を操作する例として、ユーザが指定するある文書と類似度の高い特徴次元の削除を行う例を示す。

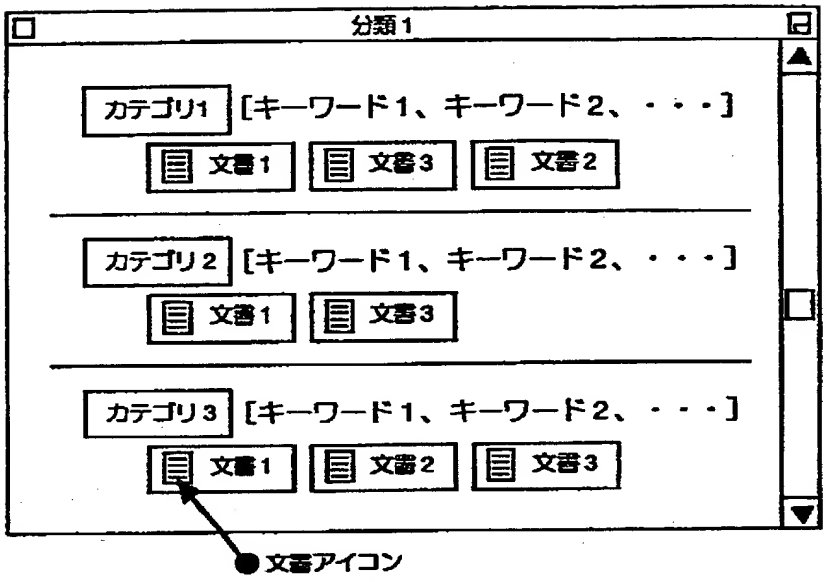
分割文書３

ユーザにより指定された文書を前記文書特徴ベクトルと同じ次元数をもつベクトルで表現し、その文書ベクトルに前記表現空間変換関数を適用し文書ベクトルを前記表現空間変換関数により構成される空間へ射影する。そして、この射影された文書ベクトルと各特徴次元との類似度を算出することで、類似度の高い特徴次元を判別する。このとき、類似度を測るための尺度としては、余弦尺度、内積尺度、ユークリッド距離尺度などを用いることができる。

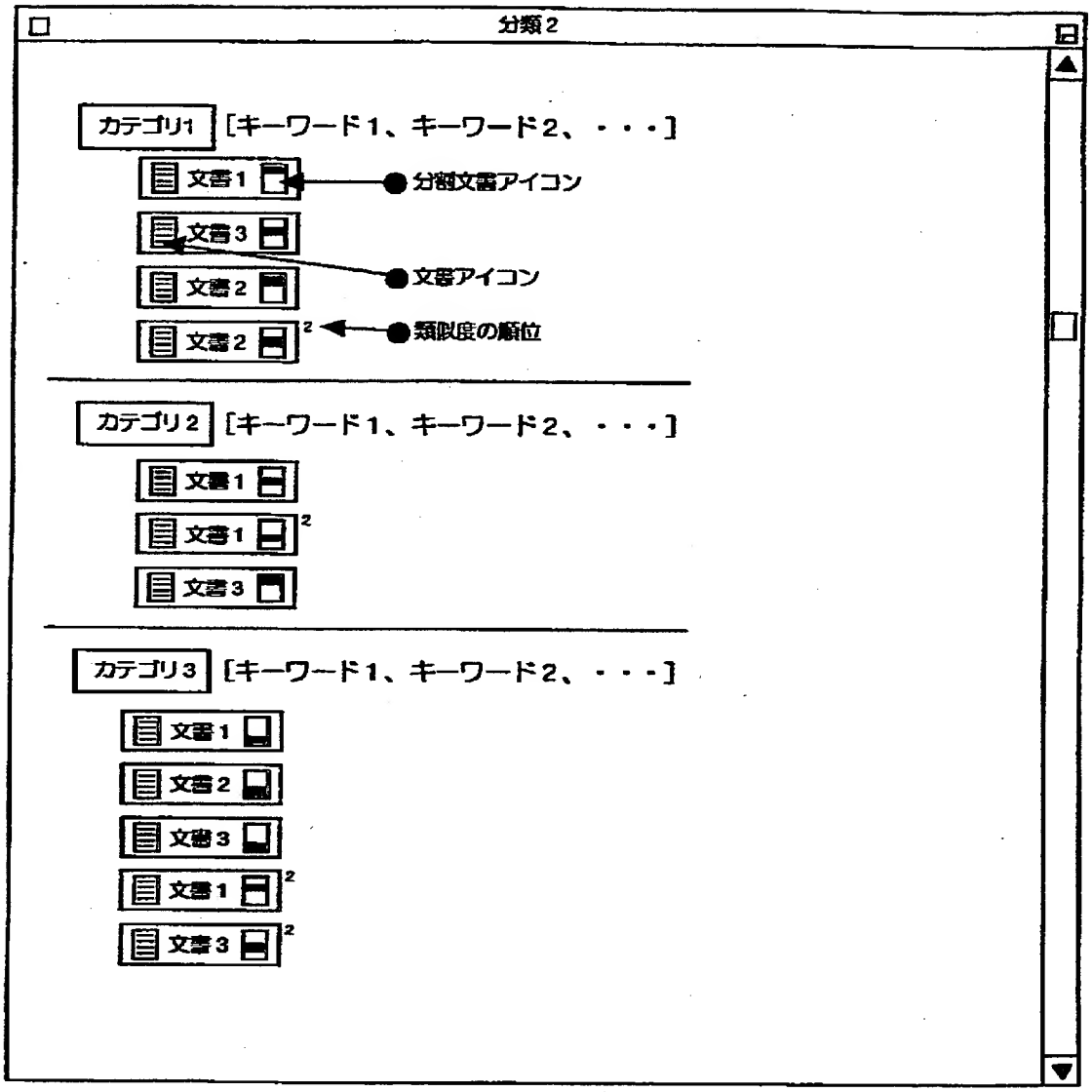
分割文書４

また、判別に関しては、ある類似度以上を削除対象として採用するような閾値処理よる判別、類似度の高い順にある一定数を削除対象として採用する定数処理、もしくは判別分析なども用いることができる。このようにして、採用された特徴次元を前記表現空間変換関数から削除することで前記表現空間変換関数を修正することができる。

【図１４】


分類1
カテゴリ1
[キーワード1、キーワード2、・・・]
文書1
文書3
文書2
カテゴリ2
[キーワード1、キーワード2、・・・]
文書1
文書3
カテゴリ3
[キーワード1、キーワード2、・・・]
文書1
文書2
文書3
文書アイコン

【図１５】


分類2
カテゴリ1 [キーワード1、キーワード2、・・・]
文書1
分割文書アイコン
文書3
文書2
文書アイコン
文書2 2
類似度の順位
カテゴリ2 [キーワード1、キーワード2、・・・]
文書1
文書1 2
文書3
カテゴリ3 [キーワード1、キーワード2、・・・]
文書1
文書2
文書3
文書1 2
文書3 2

【書類名】　　要約書

【要約】

【課題】　一文書中に複数の話題などが含まれていても、特定の話題などに限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されたりせず、利用者がその分類カテゴリを良く理解できる文書分類装置などを提供する。

【解決手段】　文書群を分類する文書分類装置において、文書データ群を入力する文書入力部１、入力された文書データ群の各文書を分割し、一つの文書データから一つまたは複数の分割文書データを生成する文書分割部２、上記文書データと分割文書データとの対応を示す文書－分割文書対応マップを生成する文書－分割文書対応マップ生成部３、上記分割文書データを分類する分割文書分類部４、分類結果に基づいて分割文書分類結果情報を生成する分割文書分類結果生成部５、上記文書－分割文書対応マップと分割文書分類結果情報から上記文書データの分類結果情報を生成する文書分類結果生成部６を備えた。

【選択図】　　図１

出　願　人　履　歴　情　報

識別番号　　〔０００００６７４７〕

１．変更年月日　　１９９０年　８月２４日
［変更理由］　新規登録
住　所　東京都大田区中馬込１丁目３番６号
氏　名　株式会社リコー